been encoded via a biharmonic potential, with folding forced to proceed sequentially via successive implementation of those constraints. Using such methods, those authors reported that a substantial number of tertiary constraints were required to assemble a three-dimensional protein structure.

Another effort to predict the global fold of a protein from a limited number of distance constraints has been reported by Aszodi *et al.*[5] Their approach was based on the use of distance geometry where a set of experimentally derived tertiary distance constraints was supplemented by a set of predicted distances between amino acid residues. The predicted distances were reported as being obtained from patterns of conserved hydrophobic amino acid residues that had been extracted from multiple sequence alignments with respect to the parent sequence. In general, they reported that when assembling structures below 5 Å cRMSD, on average, more than N/4 constraints are required, where "N" is the number of amino acid residues in the protein. Even then, the method reported by Aszodi *et al.* had difficulty selecting the correct fold from competing alternatives, although the approach was very rapid, with a calculation taking on the order of minutes on a typical contemporary workstation.

SUMMARY OF THE INVENTION

It is the object of this invention to provide new algorithms, methods, and computer systems that employ partial knowledge of known protein secondary structure and a small number of tertiary, or long range, constraints to determine the three-dimensional structure of a "target" protein. Here, "partial knowledge" means that there is no requirement for a detailed description of the local secondary structure in terms of ϕ and Ψ bond angles or their lattice equivalent. Instead, a three-letter code for secondary structure (H-helix, E-extended, and (-) everything else) is used as an input, wherein each amino acid residue of the protein is assigned an H, E, or - code. For a given protein, its corresponding H/E/- code is translated by software into loosely defined preferred ranges of local intrachain distances. It is not necessary that all, or even a part, of the three-dimensional structure of the target

protein be known, as the invention can be practiced using primary amino sequence information, whether derived from protein sequencing experiments or deduced from the coding region of a nucleic acid encoding the protein.

In particular, the invention relates to a new lattice protein model, termed "SICHO" (Side Chain Only), that focuses explicitly on the side chain center of mass positions of the amino acid residues of a target protein, and treats the protein backbone. The force field used in SICHO comprises short-range interactions that reflect secondary propensities and short-range packing biases, a geometrically implicit model of cooperative hydrogen bonds, and explicit burial, that is residues buried in the protein core and not exposed to water, pair interactions between side chains, and multi-body, involving three or more side chains tertiary interactions. The advantages afforded by the invention are due to more efficient protein representations and a new definition of the model force field that, when combined with a small number of long-range harmonic constraints (*e.g.*, known side chain contacts), result in rapid collapse and assembly of a three-dimensional structure of the target protein. Additionally, because of the way the model and force field are implemented, SICHO's computational efficiency scales with a lower portion of the chain length, *i.e.*, the number of amino acid residues comprising the target protein. Accordingly, the invention provides for the rapid, computationally efficient generation of one or more three-dimensional structures of one or more target proteins of known or deduced amino acid sequence.

Thus, a first aspect of the invention concerns methods for converting an alignment of a probe or "target" amino acid sequence with a template amino acid sequence into one or more three-dimensional reduced protein models comprising representations of side chains of amino acid residues comprising the target amino acid sequence. In some embodiments, the target amino acid sequence comprises a sequence of all of the amino acid residues of a protein. In other embodiments, the target amino acid sequence comprises a sequence of less than all of the amino acid residues of a protein, for example, a protein fragment or protein domain. A "probe

amino acid sequence" is a sequence of amino acid residues whose three-dimensional structure or a "target amino acid sequence"is being determined byt hemethods of the invention, and can also be referred to as a "target" amino acid sequence, protein, protein fragment, or domain. In some embodiments of the invention, the target amino acid sequence will be deduced from a nucleotide sequence.

A "template" amino acid sequence refers to a sequence of amino acid residues against which the target amino acid sequence is comparatively aligned. Typically, the template amino acid sequence, in addition to having a known sequence of amino acid residues, will also comprise structural or conformation information. For example, such information can include secondary, supersecondary, tertiary, or quaternary structural information.

Target and template amino acid sequences can be aligned by any suitable method. Representative alignment algorithms are described below, and any suitable alignment algorithm can be employed in the practice of the invention. In preferred embodiments, the alignment is a threading alignment, prepared by a threading algorithm.

In various embodiments, the conversion of an alignment of a target amino acid sequence with a template amino acid sequence into one or more three-dimensional reduced protein models comprising representations of side chains of amino acid residues comprising the target amino acid sequence is performed using a computer. The alignment is input into the computer (for example, from a data storage device, another computer, a user interface, *etc.*), and a program , or computer control logic, instructs the computer (typically the processor, one or more which may be present depending on the computer used) to manipulate the alignment to produce a three-dimensional reduced protein model. Preferably, several different models are produced from any given alignment by varying one or more of the constraints imposed by the program. Each of the models can be output from the computer to an output device, *e.g.*, a projection system (for example, a monitor) or to another device, such as a storage device. Preferably, the lowest energy model, or